Techno.COM, Vol. 14, No. 4, November 2015: 299-314

ANALISIS SENTIMENT PADA SOSIAL MEDIA TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER TERHADAP KATA KUNCI "KURIKULUM 2013"

Dyarsa Singgih Pamungkas[1], Noor Ageng Setiyanto[2] , Erlin Dolphina[3]
[1,2,3]Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jalan Nakula 1 No 5-7, Semarang, 50131, (024) 3517261
E-mail : 111201106186@mhs.dinus.ac.id[1], nasetiyanto@gmail.com[2] , eerlindolphina@dsn.dinus.ac.id[3]

Abstrak

Twitter salah satu situs sosial media yang memungkinkan penggunanya untuk menulis tentang berbagai hal yang terjadi dalam sehari-hari. Banyak pengguna mentweet sebuah produk atau layanan yang mereka gunakan. Tweet tersebut dapat digunakan sebagai sumber data untuk menilai sentimen pada Twitter. Pengguna sering menggunakan singkatan kata dan ejaan kata yang salah, dimana dapat menyulitkan fitur yang diambil serta mengurangi ketepatan klasifikasi. Dalam penelitian ini menggunakan Twitter Search API untuk mengambil data dari twitter, penulis menerapkan proses n-gram karakter untuk seleksi fitur serta menggunakan algoritma Naive Bayes Classifier untuk mengklasifikasi sentimen secara otomatis. Penulis menggunakan 3300 data tweet tentang sentimen kepada kata kunci "kurikulum 2013". Data tersebut diklasifikasi secara manual dan dibagi kedalam masing-masing 1000 data untuk sentimen positif, negatif dan netral. Untuk proses latih di gunakan 3000 data tweet dan 1000 tweet tiap kategori sentimentnya. Hasil penelitian ini menghasilkan sebuah sistem yang dapat mengklasifikasi sentimen secara otomatis dengan hasil pengujian 3000 data latih dan 100 tweet data ujicoba mencapai 91 %.

Kata kunci : *Twitter, Twitter Search API, sosial media, tweet, analisis sentimen, sentimen, N-gram, Naive Bayes Classifier.*

Abstract

Twitter is one social media site that allows users to write about things that happen in everyday. Many users tweeted a product or services they use. Tweets can be used as a data source for assessing the sentiment on Twitter. Users often use abbreviations and spelling words wrong, which can complicate the features are taken and reduce the accuracy of the classification. In this study use Twitter Search API to retrieve data from twitter, we apply the n-gram characters for feature selection and use Naive Bayes classifier algorithm for automatically classifying sentiment. We uses 3300 tweet data about sentiment to keywords "kurikulum 2013". Such data manually classified and divided into each 1000 data for sentiment positive, negative and neutral. For the process of training use 3000 tweet data and 1000 tweet each sentiment category. Results of this study produce a system that can automatically classify sentiment with the results of 3000 traning data and 100 testing tweets data reaches 91%.

Keywords: *Twitter, Twitter Search API, social media, tweet, sentimen analysist, sentiment, N-gram, Naive Bayes classifier.*

1. PENDAHULUAN

Pada era sekarang merupakan zaman modern yang menjadikan internet sebagai hal wajar, masyarakat dunia sekarang ini gemar bermain social media yang merupakan bagian dari internet. Twitter, Facebook, Path, Instagram merupakan salah satu dari social media tersebut. Social media merupakan media komunikasi terbuka dan tak terbatas disana masyarakat dapat secara bebas mengemukakan pendapat mereka.

Pengguna internet di Indonesia pada akhir tahun 2013 mencapai hingga 71,19 juta orang menggunakannya [1]. Indonesia adalah negara yang memiliki pengguna sosial media yang paling aktif di asia. Indonesia memiliki 79,7 % user aktif di social media mengalahkan Filipina 78 %, Malaysia 72 %, Cina 67% [2]. Pada November 2013 Twitter memiliki 19,5 juta pengguna di Indonesia dari total 500 juta pengguna global. Twitter menjadi salah satu jejaring sosial paling besar di dunia sehingga mampu meraup keuntungan mencapai USD 145 juta [3].

Twitter merupakan social media yang dibuat oleh Jack Dorsey pada tahun 2006. Pada tahun 2013 Berdasarkan press-release Twitter ada 500 juta tweet atau kicauan oleh pengguna twitter per harinya [4]. Sebanyak 500 juta tweet tersebut akan percuma bila tidak dimanfaatkan padahal di sana ada berbagai macam opini atau pendapat tentang tentang film, selebriti, politisi, produk, perusahaan, saham. dan peristiwa yang dapat diolah menjadi bahan referensi market atau penilain terhadap sosok selebriti, tokoh,atau politisi kedepannya.

Kurikulum 2013 adalah kurikulum baru yang diterapkan pada tahun 2014 oleh pemerintah. Kurikulum ini berlaku bagi pendidikan dasar hingga menengah. Pada awal penerapan kurikulum ini banyak sekali masyarakat yang meragukan keefektifan kurikulum 2013. Kurikulum 2013 ini banyak sekali komentar-komentar yang bermunculan mulai dari pelajar, pengajar, dan orang tua siswa. Banyak sekali komentar tersebut bermunculan di social media khususnya twitter. Komentar tersebut dapat berupa opini positif maupun opini negatif. Untuk mengetahuinya maka opini-opini yang ada di twitter harus diolah untuk mengklasifikasikan opini-opini tersebut menjadi opini positif, atau opini negatif. Dengan menggunakan algoritma pengklasifikasian maka opini-opini tersebut dapat terklasifikan.

Naive Bayes classifier merupakan salah satu machine learning yang merupakan algoritma untuk mengklasifikasikan sebuah data. Naive Bayes classifier ini merupakan yang paling sesuai dengan model classifier probabilistik.

Berdasarkan latar belakang diatas, penulis memimplementasikan Naive Bayes Classifier pada Analisis sentiment pada sosial media twitter kata kunci “kurikulum 2013”.

2. METODE

2.1 Tinjauan Studi

Berdasarkan penelitian yang dilakukan oleh peneliti terdahulu, hasil penelitian tentang opinion mining menunjukan berbagai pandangan khususnya yang menggunakan metode Naive Bayes classifier. Dibawah ini merupakan hasil dari penelitian yang pernah dilakukan yang relevan dengan penelitian ini, yaitu:

Penelitian mengenai klasifikasi sentimen telah dilakukan. Pada paper [5]. Sentiment Analysis atau opinion mining adalah studi komputasional dari

opini-opini orang, appraisal dan emosi melalui entitas, event dan atribut yang dimiliki. Tugas dasar dalam analisis sentimen adalah mengelompokkan polaritas dari teks yang ada dalam dokumen, kalimat, atau fitur/ tingkat aspek apakah pendapat yang dikemukakan dalam dokumen, kalimat atau fitur entitas atau aspek bersifat positif, negatif atau netral.

Pada hasil eksperimen untuk kategorisasi teks berbahasa Indonesia didapatkan bahwa Support Vector Machine menunjukkan performansi yang sedikit lebih baik dengan akurasi 92,5% dibandingkan metode Naive Bayes classifier dengan akurasi 90% padahal metode Naive Bayes classifier adalah metode yang jauh lebih konvensional dan lebih sederhana. Sehingga pada penelitian ini ingin diketahui metode yang mana memiliki performansi yang lebih baik untuk diimplementasikan dalam sentiment analysis opini berbahasa Inggris dan berbahasa Indonesia. Sedangkan pada Paper ini [6]. Mengatakan Twitter merupakan sebuah indikator yang baik untuk memberikan pengaruh dalam penelitian. Namun masih belum banyak aplikasi dan metode analisa sentimen yang dikembangkan untuk bahasa Indonesia. Faktor-faktor keuntungan tersebut mendorong perlunya dilakukan penelitian analisis sentimen terhadap dokumen berbahasa Indonesia. Penelitian analisis sentimen ini dilakukan untuk mengetahui sentimen publik mengenai sesuatu dengan menggunakan pendekatan dalam machine learning yang dikenal dengan nama Support Vector Machine dan Maximum Entropy Part of Speech Tagging yang dikhususkan pada dokumen teks berbahasa Indonesia dengan fitur unigram.

Menurut Paper ini [7], Naive Bayes classifier dapat ditingkatkan untuk mencocokkan akurasi klasifikasi model yang lebih rumit untuk analisis sentimen dengan memilih jenis yang tepat dari fitur dan menghilangkan noise dengan pemilihan fitur yang sesuai. Naive Bayes classifier dipilih karena mereka sangat cepat untuk melatih dan dapat digunakan dengan dataset yang lebih bear. Mereka juga kuat terhadap gangguan dan kurang rentan terhadap overfitting. Kemudahan implementasi juga keuntungan besar dari Naive Bayes classifier. Menurut Paper [8], melakukan klasifikasi sentimen terhadap review film dengan menggunakan berbagai teknik pembelajaran mesin.

Teknik pembelajaran mesin yang digunakan yaitu Naïve Bayes, Maximum Entropy, dan Support Vector Machines (SVM). Pada penelitian itu juga digunakan beberapa pendekatan untuk melakukan ekstraksi fitur, yaitu unigram, unigram+bigram, unigram dan Part of Speech(POS),adjective,dan ngram+posisi. Hasil dari eksperimen yang dilakukan dipenelitian ini menemukan bahwa SVM menjadi metode terbaik ketika dikombinasikan dengan unigram dengan akurasi 82.9% .

2.2. Metode Yang di Usulkan

Naive Bayes classifier merupakan metode classifier yang berdasarkan probabilitas dan Teorema Bayesian dengan asumsi bahwa setiap variabel X bersifat bebas. [9]

1.Teori Bayesian

a. X adalah data sampel dengan kelas (label) yang tidak diketahui.

b. H merupakan hipotesa bahwa X adalah data dengan kelas (label) C. P(H) adalah peluang dari hipotesa H.

c. P(X) adalah peluang data sampel yang diamati.

d. P(X|H) adalah peluang data sampel X, bila diasumsikan bahwa hipotesa benar (valid).

e. Untuk masalah klasifikasi, yang dihitung adalah P(H|X), yaitu peluang bahwa hipotesa benar (valid) untuk data sample X yang diamati:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Naïve Bayesian Classifier mengansumsikan bahwa keberadaan sebuah atribut (variabel) tidak ada kaitannya dengan beradaan atribut (variabel) yang lain.

Karena atribut tidak saling terkait maka :

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \quad (2)$$

Bila P(X|Ci) dapat diketahui melalui perhitungan diatas maka label dari data sampel X adalah label yang memiliki P(X|Ci) * P(Ci) maksimum.

2. Kelebihan Naive Bayes classifier

a. Mudah diimplementasi

b. Memberikan hasil yang baik untuk banyak kasus

3. Kekurangan Naive Bayes classifier

a. Harus mengasumsi bahwa antar fitur tidak terkait (independent) Dalam realita, keterkaitan itu ada

b. Keterkaitan tersebut tidak dapat dimodelkan oleh Naïve Bayesian Classifier

4. Implementasi Naive Bayes classifier pada text

a. Perkirakan P probabilitas (c) masing-masing kelas c ∈ C dengan membagi jumlah kata dalam dokumen di c dengan jumlah total kata dalam korpus.

b. Perkirakan P distribusi probabilitas (w | c) untuk semua kata w dan kelas c. Hal ini dapat dilakukan dengan membagi jumlah token dari w dalam dokumen di c dengan jumlah total kata dalam c.

c. Untuk mencetak dokumen d untuk kelas c, hitung:

$$score(d,c) = P(c) * \prod_{i=1}^{n} P(w_i|c) \quad (3)$$

d. Jika hanya ingin untuk memprediksi label kelas yang paling mungkin, hanya dapat memilih c dengan nilai skor tertinggi. Untuk mendapatkan distribusi probabilitas, hitung :

$$P(c|d) = \frac{score(d,c)}{\sum_{c' \in C} score(d,c)} \quad (4)$$

Langkah terakhir adalah penting namun sering diabaikan. Model memprediksi distribusi penuh atas kelas. Dimana tugas ini adalah untuk memprediksi label tunggal, satu memilih label dengan probabilitas tertinggi. Harus diakui, meskipun, bahwa ini berarti kehilangan banyak struktur. Sebagai kelemahan model Naive Bayes adalah bahwa hal itu mengasumsikan setiap fitur untuk menjadi independen dari semua fitur lainnya. Ini adalah "naive" asumsi terlihat dalam kelipatan P (wi|c) dalam definisi skor. Jadi, misalnya, jika Anda memiliki fitur terbaik dan lain dunia terbaik, maka probabilitas mereka akan berlipat seolah-olah independen, meskipun keduanya tumpang tindih. Masalah yang sama muncul kata-kata yang sangat berhubungan dengan kata lain. [10].

5. Pengumpulan Data

Dalam penelitan ini menggunakan 3 macam data yaitu data Tweet, data kata stopword dan data kata dasar.

a. Tweet

Data Tweet di peroleh dari Search API yang di sediakan oleh Twitter, kemudian data dari API tersebut di disimpan pada database. Pada saat pengumpulan data penelitian ini peneliti memasukan keyword kurikulum 2013 untuk mendapat tweet tentang opini pada objek penelitian tersebut.

b. Stopword

Data awal stopword berdasarkan dari paper [11]. Dimana datanya berjumlah 758 kata dan di simpan di dalam database

c. Kata Dasar

Data kata dasar di dapat dari kamus bahasa Indonesia Online dimana data

kata dasar berjumlah 28526 kata kemudian data kata dasar tersebut disimpan pada database.

6. Analisa Sistem

Penelitian ini memiliki tahapan proses yaitu yang pertama adalah tahap latih yang merupakan tahap klasifikasi terhadapa tweet yang diberikan sentiment, tujuannya untuk mencari kata kunci dengan probabilitasnya yang digunakan pada proses ujicoba. Kemudian tahap selanjutnya adalah tahap ujicoba merupakan proses mengklasifikasi tweet yang belum diketahui sentiment.

Pada tahap latih yang dilakukan adalah sebagai berikut:

a.Memasukkan data latih yang telah diberikan sentimentnya
b.Kemudian dilakukan proses textprocessing dan filtering
c.Setelah melakukan textprocessing dan filtering cari data n-gramnya. Data n-gram yang dicari dibandingkan dengan data n-gram yang ada didalam database
d. Jika ada maka tambahkan frekuensi katanya jika belum ada maka kata tersebut jadikan kata baru dan tambahkan frekuensi kata n-gramnya
e. Hitung probabilitas setiap n-gram P(xi|vj)
f. Ulangi langkah nomor 4 hingga 6 sampai data terdokumentasikan.
g. Tambahkan jumlah frekuensi dokumen.
h. Hitung probabilitas dokumen tweet setiap kategori sentimen P(Vj)
i. Hasilnya adalah nilai kemungkinan setiap kata n-gramnya dan nilai probabilitas setiap sentiment.
j. Proses latih selesai.

Pada tahap ujicoba yang dilakukan adalah sebagai berikut:
a. Masukan Data tweet mentah.
b. Sistem melakukan textprocessing dan filtering
c. Kata hasil textprocessing dan filtering dicari n-gram katanya, n-gram yang muncul dibandingkan dengan n-gram yang ada di database.
d. Jika ada maka nilai probabilitas yang ada di tabel pengetahuan pada database menjadi probabilitas kata, jika tidak ketemu maka frekuensi kemunculan n-gramnya bernilai 0(nol) maka hitung nilai probabilitas tiap n-gram-nya.
e. Ulangi langkah nomor 3 hingga 5 sampai data katterdokumentasikan.
f. Hitung nilai probabilitas di setiap kategori sentiment
g. Mencari nilai probabilitas tertinggi antara sentiment positif, negatif, atau netral.
h. Tentukan sentiment tweet tersebut
i. Masukan data tweet tadi ke tabel data uji dengan ditambahkan sentimentnya.
j. Proses ujicoba selesai.

Text Preprocessing merupakan salah satu langkah dalam 2 tahap diatas, dimana text preprocessing melakukan beberapa filtering terhadap sebuah tweet berikut proses melakukan text preprocessing:
a. Tweet tersebut dijadikan huruf kecil semua disamakan semua hurufnya.
b. Lakukan filtering dengan menghapus URL, mention (misal: @IDmaju), hashtag (misal: #bisa), dan RT atau retweet.
c. Kemudian Hapus juga tanda baca dan special karakter.
d. Tersisa hanya kata-kata.
e. Proses selesai.

Setelah dilakukan proses text preprocessing kemudian hasil dari text preprocessing dilakukan pembandingan kata yang di tabel stopword pada database jika terdapat kata tersebut maka kata tersebut dihapus. Jika tidak ada maka kata tersebut tidak dihapus. Kemudian kata atau kalimat tersebut di lakukan proses stemming dengan menghilangkan imbuhan kata dan menjadikan kata tersebut kata dasar.

3. HASIL DAN IMPLEMENTASI

1. Penggunaan *Naive Bayes Classifier*
Pengunaan Naive Bayes Classifier pada subbab ini akan dibahas cara Naive Bayes Classifier mengklasifikasikan sebuah kalimat atau tweet. Berikut contoh penggunaannya :

Tabel 1 : Data pengetahuan sentimen positif

id	n-gram	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	2	0.083333
2	ur	1	0.055556
3	ri	2	0.083333
4	ik	2	0.083333
5	ul	1	0.055556
6	lu	1	0.055556
7	mu	1	0.055556
8	m_	1	0.055556
9	_2	1	0.055556
10	20	2	0.083333
11	01	2	0.083333
12	13	1	0.055556
13	3_	1	0.055556
14	_o	2	0.083333
15	ke	1	0.055556

Jumlah frekuensi keseluruhan(n) adalah 21
Jumlah n-gram = 15
Tabel 2 merupakan data pengetahuan dengan sentimen negatif

Tabel 2 : Data pengetahuan sentimen negatif

id	n-gram	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	2	0.069767
2	ur	1	0.046512
3	ri	2	0.069767
4	ik	2	0.069767
5	ul	1	0.046512
6	lu	1	0.046512
7	mu	1	0.046512
8	m_	1	0.046512
9	_2	1	0.046512
10	20	2	0.069767
11	01	2	0.069767
12	13	1	0.046512
13	3_	1	0.046512
14	_j	2	0.069767
15	je	1	0.046512
16	el	1	0.046512
17	le	1	0.046512
18	ek	2	0.069767

Jumlah frekuensi keseluruhan(n) adalah 25
Jumlah n-gram = 18
Tabel 3 merupakan data pengetahuan dengan sentimen netral

Tabel 3 : Data pengetahuan sentimen netral

id	n-gram	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	2	0.075
2	ur	1	0.05
3	ri	2	0.075
4	ik	2	0.075
5	ul	1	0.05
6	lu	1	0.05
7	mu	1	0.05
8	m_	1	0.05
9	_2	1	0.05
10	20	2	0.075
11	01	2	0.075
12	13	1	0.05

13	3_	1	0.05
14	_b	2	0.075
15	ba	1	0.05
16	ar	1	0.05
17	ru	1	0.05

Jumlah frekuensi keseluruhan(n) adalah 23
Jumlah n-gram = 17
Dari ketiga tabel, tabel 4.1, tabel 4.2, tabel 4.3 maka diperoleh nilai P(Vj):
P(positif) = 1/3 = 0,5
P(negatif) = 1/3 = 0,5
P(netral) = 1/3 = 0,5

Tabel 5 merupakan data atau tweet yang akan di klasisfikasikan

Tabel 5: Tweet yang akan di klasifikasikan

id	n-gram
1	ku
2	ur
3	ri
4	ik
5	ul
6	lu
7	mu
8	m_
9	_2
10	20
11	01
12	13
13	3_
14	_h
15	eh
16	eb
17	ba
18	at

Pada tahap klasifikasi dimulai dengan pencarian nilai probabilitas dengan membandingkan kata-kata pada tabel diatas dengan tabel data pengetahuan.

Tabel 6 : Pencarian probabilitas positif tweet yang akan di klasifikasikan

id	n-gram	n-gram positif	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	ku	2	0.083333
2	ur	ur	1	0.055556
3	ri	ri	2	0.083333
4	ik	ik	2	0.083333
5	ul	ul	1	0.055556
6	lu	lu	1	0.055556
7	mu	mu	1	0.055556
8	m_	m_	1	0.055556
9	_2	_2	1	0.055556
10	20	20	2	0.083333
11	01	01	2	0.083333
12	13	13	1	0.055556
13	3_	3_	1	0.055556
14	_h	_o	0	0.027778
15	he	ke	0	0.027778
16	eb	-	0	0.027778
17	ba	-	0	0.027778
18	at	-	0	0.027778

$$V_{map} = \underset{vj\ ev}{argmax}\ P(x_1, x_2, x_3, \ldots . x_n|v_j)P(v_j)$$

$$Vmap(positif) = 0.083333 * 0.055556 * 0.083333 * 0.083333 * 0.055556 * 0.055556 * 0.055556 * 0.055556 * 0.055556 * 0.083333 * 0.083333 * 0.055556 * 0.055556 * 0.027778 * 0.027778 * 0.027778 * 0.027778 * 0.027778 * 0.5 = 3.0158E-24$$

Tabel 7 : Pencarian probabilitas negatif tweet yang akan di klasifikasikan

id	n-gram	n-gram positif	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	ku	2	0.069767
2	ur	ur	1	0.046512

3	ri	ri	2	0.069767
4	ik	ik	2	0.069767
5	ul	ul	1	0.046512
6	lu	lu	1	0.046512
7	mu	mu	1	0.046512
8	m_	m_	1	0.046512
9	_2	_2	1	0.046512
10	20	20	2	0.069767
11	01	01	2	0.069767
12	13	13	1	0.046512
13	3_	3_	1	0.046512
14	_h	_j	0	0.023256
15	he	je	0	0.023256
16	eb	el	0	0.023256
17	ba	le	0	0.023256
18	at	ek	0	0.023256

$$Vmap(negatif) = 0.069767 * 0.046512 * 0.069767 * 0.069767 * 0.046512 * 0.046512 * 0.046512 * 0.046512 * 0.046512 * 0.069767 * 0.069767 * 0.046512 * 0.046512 * 0.023256 * 0.023256 * 0.023256 * 0.023256 * 0.023256 * 0.5 = 1.23145E-25$$

Tabel 8 : Pencarian probabilitas netral tweet yang akan di klasifikasikan

id	n-gram	n-gram positif	frekuensi (n_k)	Probabilitas $P(x_i\|V_j)$
1	ku	ku	2	0.075
2	ur	ur	1	0.05
3	ri	ri	2	0.075
4	ik	ik	2	0.075
5	ul	ul	1	0.05
6	lu	lu	1	0.05
7	mu	mu	1	0.05
8	m_	m_	1	0.05
9	_2	_2	1	0.05
10	20	20	2	0.075
11	01	01	2	0.075
12	13	13	1	0.05
13	3_	3_	1	0.05
14	_h	_b	2	0.025
15	he	ba	0	0.025
16	eb	ar	0	0.025
17	ba	ru	2	0.075
18	at	-	0	0.025

$$Vmap(netral) = 0.075 * 0.05 * 0.075 * 0.075 * 0.05 * 0.05 * 0.05 * 0.05 * 0.05 * 0.075 * 0.075 * 0.05 * 0.05 * 0.025 * 0.025 * 0.025 * 0.075 * 0.025 * 0.5 = 1.35787E-24$$

Pada hasil perhitungan nilai Vmap dari tweet yang akan diklasifikasikan nilai Vmap positif lebih tinggi dari Vmap negatif dan Vmap positif sehingga tweet tersebut masuk dalam kategori sentimen positif

3. Gambaran Umum Sistem

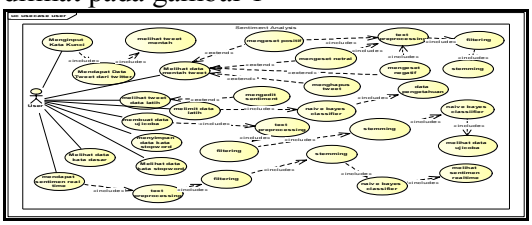
Sistem Website ini merupakan sistem machine learning yang menganalisis sentimen pada kata kunci yang di berikan oleh user. User harus menginputkan kata kunci yang ingin di analisis kemudian sistem akan menampilkan sentimen pada kata kunci tersebut. Sebelum sistem telah di berikan dapat menganalisa kata kunci secara real-time sistem sudah diberikan data latih dan data ujicoba sebagai data pengetahuan untuk menganalisa kata kunci yang diberikan oleh user. Terdapat 1 aktor yang dalam sistem ini yaitu user. User melakukan pencarian analisis sentimen dengan menginputkan kata kunci, dan juga bertugas membuat data latih dan data ujicoba sebagai data pengetahuan.

a. Perancangan Sistem

Perancangan sistem aplikasi ini menggunakan beberapa macam diagram agar mudah untuk membangun aplikasi diantaranya menggunakan Use case diagram, dan sequence diagram berikut diagram tersebut:

1. Diagram Use case

Gambaran diagram Use case user dapat dilihat pada gambar 1



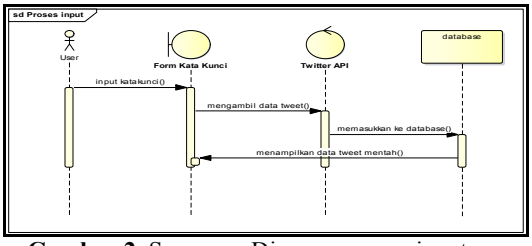
Gambar 1. Use case user

2. Sequence Diagram

Berikut adalah sequence diagram dari beberapa proses yang ada pada sistem yaitu sebagai berikut:

a. Diagram Sequence Proses input kata kunci

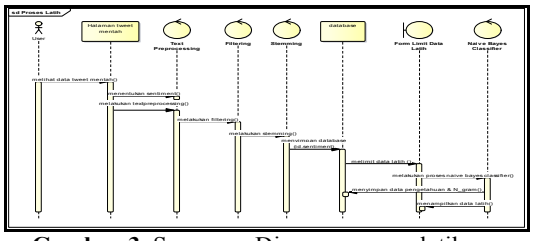
Diagram Sequence proses input kata kunci merupakan proses yang harus di lakukan user. User menginputkan kata kunci kemudian website akan mengambil data melalui Twitter API kemudian di tampilkan dan juga di masukkan ke database.



Gambar 2. Sequence Diagram proses input kata kunci

b. Diagram Sequence Proses Latih

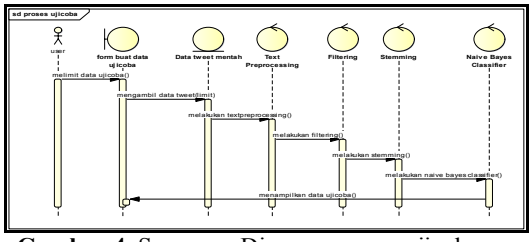
Diagram Sequence proses latih merupakan proses user melakukan sentimen pada tweet mentah kemudian tweet tersebut di proses textpreprocessing, filtering, dan stemming setelah itu user harus memnbuat data pengetahuan dengan cara melimit data latih per sentimen kemudian dilakukan proses Naive Bayes classifier yang akan disimpan di database dan ditampilkan di halaman limit data latih



Gambar 3. Sequence Diagram proses latih

c. Diagram Sequence Proses Ujicoba

Diagram Sequence proses ujicoba merupakan proses user melakukan ujicoba pada data latih yang di jadikan data pengetahuan untuk membentuk atau menentukan sentimen pada tweet mentah.



Gambar 4. Sequence Diagram proses ujicoba

3. Perancangan Database

Data tweet yang didapat dimasukkan ke database untuk diolah, data tersebut adalah data latih, data n_gram, data kata stopword, dan data kata dasar.

a. Tabel rawtweet

Tabel ini berisi data hasil mengambil dari twitter *Search API*.

Tabel 9: Perancangan Database Tabel Rawtweet

No	Nama Field	Tipe	Deskripsi
1	id	int(20)	id dari tweet sebagai *Primary Key*
2	tweet_id	bigint(20)	id dari yang diberikan twitter pada setiap tweet

3	tweet_text	varchar(160)	tweet dari setiap user
4	created_at	datetime	tanggal dari tweet
5	user_id	bigint(20)	id dari user twitter
6	screen_name	char(20)	screen name setiap user twitter
7	profile_image_url	varchar	Url dari foto user twitter

b. Tabel datalatih

Tabel ini merupakan database tweet yang sudah diberikan sentimen secara manual, tabel ini akan diolah untuk dijadikan data pengetahuan.

Tabel 10: Perancangan Database Tabel datalatih

No	Nama Field	Tipe	Deskripsi
1	tweet_id	bigint(20)	id dari yang diberikan twitter pada setiap tweet sebagai *Primary Key*
2	tweet_text	varchar(160)	tweet dari setiap user
3	screen_name	char(20)	screen name setiap user twitter
4	sentimen	varchar(10)	sentimen dari setiap tweet yang diberikan

c. Tabel tb_stopword

Tabel ini merupakan database kata stopword yang digunakan pada tahap filtering. Kata stopword adalah kata umum (common words) yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna.

Tabel 11: Perancangan Database Tabel Kata Stopword

No	Nama Field	Tipe	Deskripsi
1	id_stopword	int(10)	id dari setiap kata stopword sebagai *Primary Key*
2	katastopword	varchar(70)	kata stopword

d. Tabel tb_Dasar

Tabel ini merupakan database kata dasar yang digunakan pada tahap stemming. Stemming merupakan proses membuat sebuah kata berimbuhan menjadi kata dasar.

Tabel 12: Perancangan Database Tabel Kata Dasar

No	Nama Field	Tipe	Deskripsi
1	id_katadasar	int(10)	id dari setiap kata dasar sebagai *Primary Key*
2	katadasar	varchar(70)	kata dasar
3	tipe_katadasar	varchar(25)	tipe kata dasar

e. Tabel n_gram

Tabel ini merupakan database hasil peng-ngram-an dari data latih yang dijadikan data pengetahuan.

Tabel 13: Perancangan Database Tabel N_gram

No	Nama Field	Tipe	Deskripsi
1	kd_ngram	int(3)	id dari setiap n_gram sebagai

			Primary Key
2	n_gram	varchar(3)	kata N_gram
3	sentimen	varchar(10)	sentimen dari setiap n_gram
4	frekuensi	float	frekuensi kemunculan n_gram setiap kategori sentimen
5	probabilitas	float	probabilitas setiap n_gram pada data pengetahuan atau setiap dokumen atau tweet

f. Tabel n_gramsementara

Tabel ini merupakan database hasil peng-ngram-an sementara dari data latih yang sebelum di masukkan ke tabel n_gram.

Tabel 14: Perancangan Database Tabel n_gram sementara

No	Nama Field	Tipe	Deskripsi
1	kd_ngram	int(3)	id dari setiap n_gram
2	n_gram	varchar(3)	kata N_gram
3	sentimen	varchar(10)	sentimen dari setiap n_gram
4	frekuensi	float	frekuensi kemunculan n_gram setiap kategori sentimen

g. Tabel Dok_sementara

Tabel ini merupakan database untuk menyimpan banyaknya jumlah dokumen atau tweet yang digunakkan pada data pengetahuan.

Tabel 15: Perancangan Database Tabel dok_sementara

No	Nama Field	Tipe	Deskripsi
1	probabilitas	double	nilai probabilitas setiap kategori sentimen
2	sentimen	varchar(10)	jenis sentimen
3	jumdok	int(10)	jumlah dokumen yang digunakan pada data pengetahuan
4	jumngram	int(10)	jumlah n-gram keseluruhan yang digunakan pada data pengetahuan

4. Implementasi

Pengembangan ini bertujuan untuk mengetahui sudah sejauh mana kemajuan dalam pembuatan website analisis sentimen menggunakan algoritma naive bayes classifier dalam pengembangannya, sehingga dapat dilakukan perubahan atau perbaikan jika terdapat masukan dari pemakai.

a. Halaman Utama

Halaman utama merupakan halaman yang akan pertama kali muncul. Pada halaman utama terdapat menu navigasi dan menu tombol.

Gambar 15. Halaman Utama Website

b. Tambah Tweet

Pada Menu ini terdapat form kata kunci yang di gunakan untuk mencari tweet yang behubungan dengan kata kunci yang kemudian di simpan di database dan di tampilkan ke layar.

Gambar 16. Halaman Tambah Tweet

c. Lihat Tweet

Menu Lihat tweet terdapat tabel data tweet mentah yang akan disentimenkan, terdapat 4 tombol pilihan yaitu set to positif, set to negatif, set to netral, dan delete.

Gambar 17. Halaman Lihat Tweet

d. Data Latih

1. Lihat Data Latih

Menu lihat data latih terdapat tabel data latih yang telah di buat pada menu lihat tweet, pada menu lihat data latih dapat juga mengedit sentimen yang telah di lakukan dengan cara mengklik tombol edit kemudian muncul modal atau pop-up pilihan tombol.

(a)

(b)

Gambar 18. (a) Halaman data latih (b) Pop-up modal edit sentimen

2. Buat Data Latih

Menu buat data latih terdapat form yang berfungsi melimit data latih yang akan di jadikan data pengetahuan dengan cara mengklik tombol buat data latih yang akan melimit jumlah data latih setiap kategori sentimen yang akan dijadikan data pengetahuan kemudian memproses tweet tersebut menjadi n-gram, setiap n-gram menyimpan sentimen dari masing-masing tweet dan setiap n-gram memiliki nilai probabilitas pada database, kemudian akan menyimpan banyaknya data tiap kategori sentimen pada data pengetahuan.

Gambar 19. Halaman Buat Data Latih

e. Data Uji Coba

Menu data uji coba terdapat form yang berfungsi melimit tweet mentah yang akan di ujicoba di analisis sentimennya dengan cara mengklik tombol buat data uji yang akan melimit jumlah data tweet mentah kemudian akan di proses penganalisan sentimennya kemudian di tampilkan dilayar.

Gambar 20. Halaman Data Ujicoba

f. Stopword
1. Lihat Stopword
Menu lihat stopword terdapat tabel data kata stopword.

Gambar 21. Halaman Lihat Stopword

2. Tambah Stopword
Menu tambah. stopword terdapat form yang digunakan untuk menambahkan data kata stopword.

Gambar 22. Halaman Tambah Stopword

g. Kata Dasar
Menu kata dasar berisi tabel data kata dasar yang digunakan pada proses stemming.

Gambar 23. Halaman Kata Dasar

h. Real Time Sentiment
Menu data Real Time sentiment terdapat form yang berfungsi kata kunci yang akan di ujicoba di analisis sentimennya dengan cara mengklik tombol buat data uji yang mencari tweet mentah menggunakan twitter search API kemudian akan di proses penganalisan sentimennya kemudian di tampilkan dilayar.

Gambar 25. Halaman Real Time Sentiment

5. Pengujian
a. Pengujian Black Box
Pengujian ini memakai teknik black box, dimana yang diuji adalah fungsi-fungsi yang digunakan untuk membuat sebuah website analisis sentiment. Pengujian ini dilakukan dengan membuka menu-menu yang mengiakan fungsi-fungsi dari sentimen analisis. Pengujian ini juga memastikan website analisis sentimen berjalan dengan baik.

Tabel 16: Hasil pengujian dengan Black Box

No	Skenario Pengujian	Test Case	Hasil yang diharapkan	Hasil Pengujian
1	User melakukan penyentiment untuk data latih dari data tweet mentah	Load fungsi filtering, stopword dan stemming	Sistem menyimpan tweet pada data latih kata-kata sudah menjadi kata dasar dan juga sudah tidak ada kata stopword ,mention, RT dan tanda baca	Sistem bisa menyimpan tweet yang sudah di bersih dan menjadi kata dasar
2	User melakukan limitasi	Load fungsi naive	Sistem menyimpan	Sistem bisa menyimpan dan

	data latih	bayes classifier	menghitung probabilitas n-gram setiap tweet yang dijadikan data latih sejumlah sesuai dengan limitasi tweet dan juga menyimpan jumlah tweet setiap kategori sentiment pada data pengetahuan	mengihitung probabilitas n-gram setiap tweet setiap sentiment
3	User melakukan limitasi data ujicoba	Load data latih dan fungsi naive bayes classifier	Sistem menampilkan data tweet	Sistem bisa menampilkan data tweet yang sudah ditambahkan sentimentnya
4	User menginput kata kunci yang akan dicari sentimentnya	Load data latih dan fungsi filtering, stemming, dan stopword juga fungsi naive bayes classifier	Sistem menampilkan data tweet	Sistem bisa menampilkan data tweet yang sudah ditambahkan sentimentnya

b. Hasil Penelitian

Berdasarkan langkah perancangan dan implementasi yang dibuat maka terdapat beberapa hasil penelitian yang didapat selama penelitian. Pengujian telah dilakukan pada fungsi-fungsi website analisis sentimen menggunakan Black Box. Setelah di uji di dapatkan hasil pengujian dengan menguji data latih dan data ujicoba menghasilkan data sebagai berikut :

Tabel 17: Hasil Pengujian

Data Latih			Data Ujicoba	Akurasi
Positif	Negatif	Netral		
100	100	100	100	72 %
200	200	200	100	74 %
300	300	300	100	75 %
400	400	400	100	76 %
500	500	500	100	77 %
600	600	600	100	77 %
700	700	700	100	78 %
800	800	800	100	84 %
900	900	900	100	88 %
1000	1000	1000	100	91 %

4. KESIMPULAN DAN SARAN

4.1.Kesimpulan

Setelah melakukan pembangunan website analisa sentimen pada sosial media twitter menggunakan naive bayes classifier terhadap kata kunci "kurikulum 2013", maka peneliti

menyimpulkan beberapa hal, yaitu sebagai berikut :

1. Klasifikasi tweet bersentimen lebih akurat jika data latih yang di gunakan semakin banyak dalam data pengetahuan.
2.Akurasi Naive Bayes Classifier memberikan hasil sebesar 91 % untuk 1000 data latih yang diberikan.
3. Fungsi N-gram kata dapat meningkatkan analisis sentimen.
4. Jika hasil Vmap setiap kategori sama akan menghasilkan kategori tidak tersentimentkan.
5. Jika hasil Vmap pada tweet ada yang berjumlah nol(0) maka data pengetahuan kurang.
6. Analisis tidak berjalan maksimal terhadap bahasa asing dan bahasa daerah.

4.2 Saran

Untuk meningkatkan kinerja serta menyempurkan sistem yang telah dibuat maka peneliti memberikan saran sebagai berikut :

1. Pada penelitian berikutnya dapat di tambahkan fitur yang mendeteksi emoticon dan juga mengetahui posisi sebuah kata dalam kalimat menggunakan Part of Speech Tagging dalam proses pengklasifikasian.
2. Bahasa yang di gunakan tidak hanya bahasa Indonesia tetapi bisa menggunakan bahasa asing dan bahasa daerah.

DAFTAR PUSTAKA

[1] "Merdeka.com," Web Newsportal, 15 januari 2014.[Online].Available: http://www.merdeka.com/teknologi/jumlah-pengguna-internet-indonesia-capai-7119-juta-pada-2013.html. [Diakses 2014 oktober 2014].

[2] "Kementrian Komunikasi dan Informatika," Web Kementrian, 7 November 2013. [Online]. Available: http://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker. [Diakses 19 Oktober 2014].

[3] "Globalstats Research," Research, 2 Agustus 2013. [Online]. Available: http://www.globalstats-research.com/penggunaan-media-sosial-di-indonesi/. [Diakses 2014 Oktober 20].

[4] "Telegraph," Web Newsportal, 21 Maret 2013. [Online].Available: http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html. [Diakses 15 Oktober 2014].

[5] N. W. S. Saraswati, "NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINES," dalam Seminar Nasional Sistem Informasi Indonesia, 2013.

[6] N. D. Putranti dan E. Winarko, "Analisis Sentiment Twitter Untuk Teks Bahasa Indonesia dengan Maximum Entropy dan Support Vector Machine," IJCCS, vol. 8, no. 1, pp. 91-100, 2014.

[7] V. Narayanan, I. Arora dan A. Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model.".

[8] B. Pang , L. Lee dan S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning," dalam ACL-02 conference on Empirical methods in natural language processing-Vo, 2002.

[9] S. M. Dr. Taufik Fuadi Abidin, "Naiive Bayesian Classifier," FMIPA Universitas Syiah Kuala, Banda Aceh, 2013.

[10] S. L. Christopher Potts, "Sentiment Symposium Tutorial: Classifiers," 2011. [Online]. Available: http://sentiment.christopherpotts.ne

t/classifiers.html#others. [Diakses 27 Desember 2014].

[11] F. Z. Tala, "A Study of Stemming Effects on Information Retrival in Bahasa Indonesia," Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands, Amsterdam, 2003.